PHIL 7001: Homework Assignment 1

Professor Boris Babic

Instructions:

This is the first homework assignment for PHIL 7001. The assignment will be graded out of a total of 100 points. It is open notes, and open book – this means that you can use any resources you like in order to formulate your answer. However, please do not "copy" from internet materials.

Course materials: You can directly use course materials – both lectures and readings – without citation.

Non course materials: If you directly rely on external sources from the internet or other textbooks or articles, then please provide a citation to those references.

This homework assignment shall be submitted via the course website by **Thursday, October 19, 11:59pm**.

In order to write your answers you can use a word processor, or (for mathematics) you can write your answers in pen and then scan them into a pdf. If you know how to use LaTeX, then you can also use LaTeX. But if you don't know how to use it, then you don't have to. Whichever method you choose, please ensure that you submit your homework assignment as one **single pdf**. And please make sure that everything within this single PDF is clearly legible. This will make it much easier for grading purposes.

Calculators are allowed, and you can use any calculator you wish.

Please write your name and student number clearly on the first page of your submission, and please include page numbers at the bottom of each page of your submission.

There should be no trick questions on the homework. If anything seems unclear or ambiguous, then you can use your judgment to resolve the ambiguity, and note that you have done so in your answer. For example: if you are not sure whether to write something mathematically, or in words, it would be safe to do both, or to pick one option and explain why you picked it (of course: it is always best to just ask me for clarification! But if you cannot ask me, and you need to answer, then this is what I recommend).

Problem 1: Introduction to R (10 points)

We have a dataset below, named `data_students`, containing information about students' study hours and their corresponding exam scores:

```
data_students <- data.frame(
  study_hours = c(2, 4, 3, 5, 7, 6, 8, 9, 10),
  exam_scores = c(60, 75, 68, 80, 90, 85, 92, 94, 98)
)
```

You want to visualize the relationship between study hours and exam scores using a scatter plot with a regression line. The x-axis should be labeled **Study Hours**, and the y-axis should be labeled **Exam Scores**.

Please use R in order to generate a scatter plot with a regression line, and submit your code, and your plot, as your answer.

Below, I have given you a hint to make the question easier. In other words, your goal is to complete the blanks, and then execute the code in R to actually generate the plot.

HINT:

```
ggplot(_______, aes(x=_______, y=_______)) +
  geom_point() +
  geom_smooth(method = "___", se = FALSE, color = "blue") +
  labs(x='_______', y='_______',
       title='______')
```

Problem 2: Probability Basics (20 points)

a. State or describe the (three) conditions that a function, $P(A)$, must fulfill in order to be a legitimate probability function.

b. Suppose we have three events (A, B, C) and we have the following vector of illegitimate probability assignments for A, B, and C, respectively: (.6, .8, .6). Why is this vector of probability assignments illegitimate? What would you do to it in order to make it a legitimate assignment of probabilities?

c. Using the formula for the probability of a union, calculate $P(A \cup B)$ when $P(A) = 0.3$, $P(B) = 0.5$, and $P(A \cap B) = 0.1$.

d. If $P(A) = 0.6$, $P(B) = 0.8$, and $P(A \cap B) = 0.5$, what is the probability of neither event A nor event B occurring?

e. Calculate the probability of event A occurring given that event B has occurred, if $P(A) = 0.4$, $P(B) = 0.3$, and $P(A \cap B) = 0.15$.

For part (e), it will help you to know that $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Problem 3: Bayes' Rule (20 points)

Suppose that the influenza virus (i.e., the flu) this winter in Hong Kong has an occurrence rate of 0.01 (i.e., 1%). Suppose that a very reliable test has been developed to check whether or not someone has the flu.

The test has a sensitivity rate of 0.96 (i.e., 96%). This means that:
Pr(Positive Test | Patient has Disease) = 0.96.

The test has a specificity rate of 0.97 (i.e., 97%). This means that:
Pr(Negative Test | Patient does not have Disease) = 0.97.

Find the probability that a patient has the disease, given that they receive a positive test. That is, find:
Pr(Patient has Disease | Positive Test).

Problem 4: Probability Spaces (25 Points)

a. Write down the binomial distribution in mathematical form and for each term in the distribution, state whether it is a random variable, or a parameter.

b. Suppose it is known that 70% of students at HKU prefer white sneakers to black sneakers. If you ask 6 randomly selected students on campus, what is the probability that exactly 3 of them prefer white sneakers to black sneakers?

c. White down the normal distribution in mathematical form, and for each term in the distribution, state whether it is a random variable, a parameter, or a constant.

d. Suppose that a certain population follows a normal distribution with a mean of 50, and a standard deviation of 10. Using R, find the probability that a randomly selected value is greater than 65. (copy down your R code, which should be very short, as well as your final answer).

e. Suppose that a certain population follows a normal distribution with mean 10 and standard deviation of 2. Without using R, state the approximate probability that a randomly selected value is greater than 14, and explain your answer.

Problem 5: Linear Regression (25 Points)

Consider again the same data from Problem 1.

```
data_students <- data.frame(
  study_hours = c(2, 4, 3, 5, 7, 6, 8, 9, 10),
  exam_scores = c(60, 75, 68, 80, 90, 85, 92, 94, 98)
)
```

a. Using R, write a linear (regression) model, which treats the exam score as the output, and the number of hours studied as the input. In other words, you want to write a model which models exam scores as a function of study hours. Copy and paste your R code as part of your answer.

b. Using R, create a summary of the model.

c. Is the input variable (study hours) a statistically significant predictor of a student's exam score? Why or why not?

d. Mathematically, you have created the following model in part a., above:

$$\text{exam scores} = \beta_0 + \beta_1 \text{study hours} + e.$$

What is the p-value for the following null-hypothesis:

$$H_0 : \beta_0 = 0?$$

Hint: This is something you can find in your model summary.

e. Suppose that a student studies for 3 hours. What is their predicted exam score? Hint: This is something you can calculate using the model summary.